



Powering the API world

AI Gateway June 2024 Feature Rollup

Jack Tysoe

Staff Field Engineer

June 2024

AGENDA

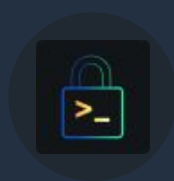
- | | |
|---|------------|
| 1. Features Right Now - Recap | 10 Minutes |
| 2. The <i>Current Problem</i> - Auditing and Safety | 15 Minutes |
| 3. Per-User Restrictions | 5 Minutes |

AI Gateway Today

Solid Foundation in Kong 3.6



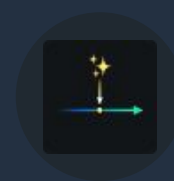
**Multiple Popular
Providers and
Models**



**Prompt Templating,
Guarding and
Decorating**



**In-Depth Contextual
Usage Statistics**



**Augment 'Normal'
APIs with LLM
Features**

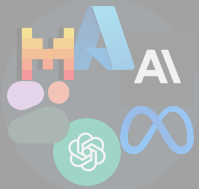
Solid Foundation in Kong 3.6



Multiple Popular
Providers and Models

**OpenAI-and-compatibles,
Llama2 and Ollama,
Mis/xtral and Ollama,
Cohere,
Anthropic**

Solid Foundation in Kong 3.6



Multiple Popular
Providers and
Models



Prompt Templating, Guarding
and Decorating

**Aid and/or restrict LLM
usage, depending on
corp requirements and/or
developer familiarity.**

Usage Statistics

Features

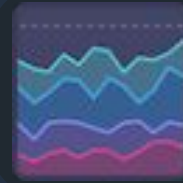
Solid Foundation in Kong 3.6



Multiple Popular
Providers and
Models

**Tokens Counts,
input/output chat
logging, latencies.**

Prompt Templating,
Guarding and
Decorating

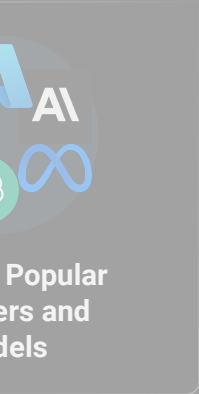


**In-Depth Contextual Usage
Statistics**




Augment 'Normal'
APIs with LLM
Features

Solid Foundation in Kong 3.6

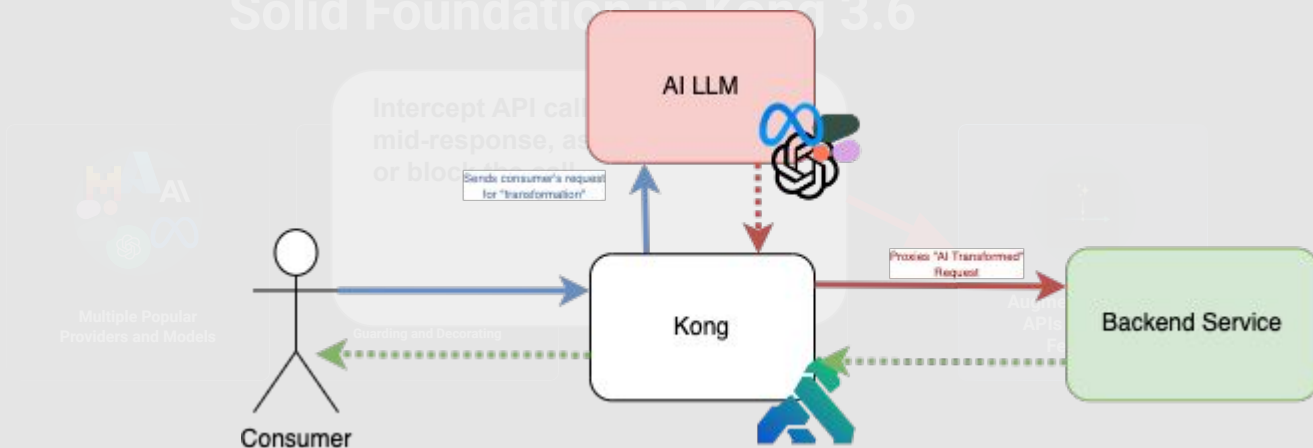


Intercept API call mid-request or mid-response, asking LLM to alter or block the call.

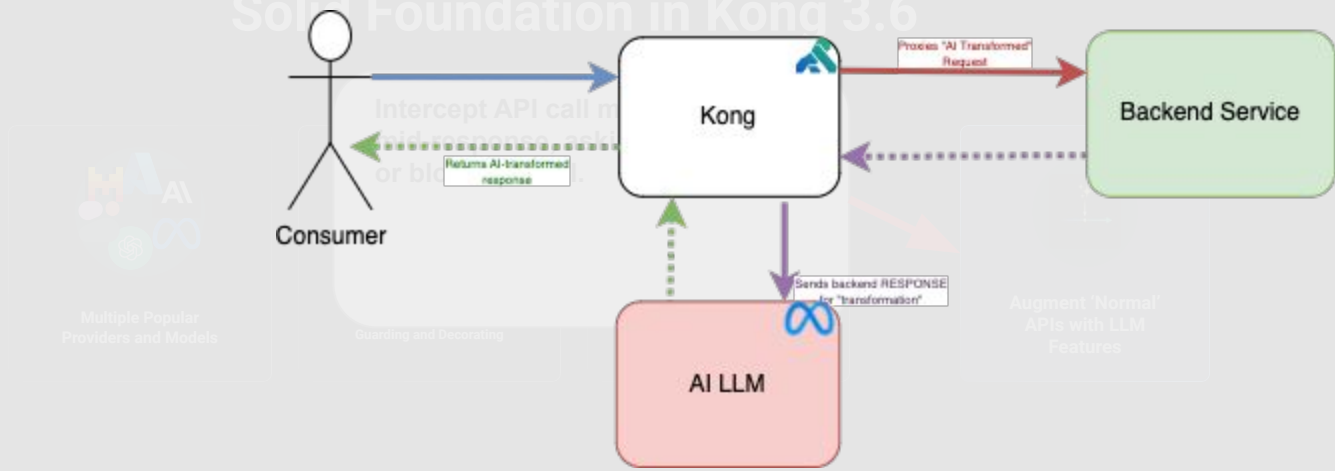


Augment 'Normal' APIs with LLM Features

Solid Foundation in Kong 3.6



Solid Foundation in Kong 3.6



Multitude of New Functionality in 3.7



**Response Streaming
and More Models**



OpenAI SDK Support



**[Enterprise]
Token-Based Rate
Limiting**



**[Enterprise] Azure
Content Safety
Introspection**

Multitude of New Functionality in 3.7



Response Streaming and More Models

Intense technical feat to capture each chunk of events, transform each into Kong-compatible format, and re-encode to client

Open

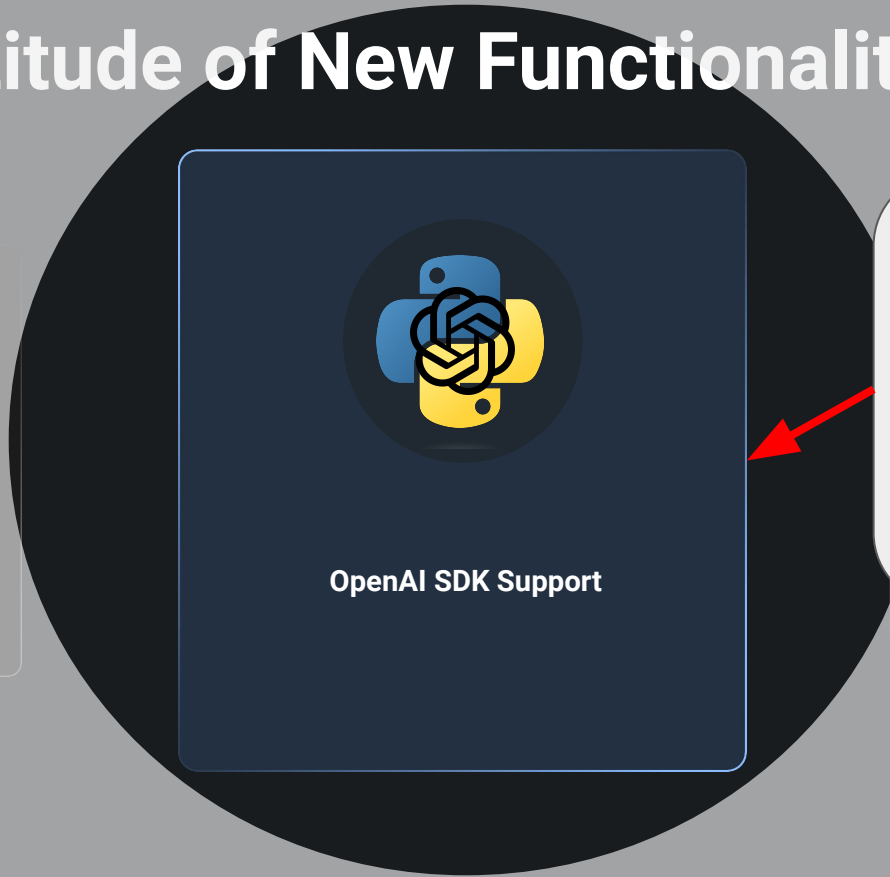
Limiting

[Enterprise
Content
Introspe

Multitude of New Functionality in 3.7



Response Streaming
and More Models



Transform/load of specific parameters, transparent to the SDK and agnostic of the selected backend model or even provider!

Multitude of New Functionality in 3.7

**Restrict
user/team/customer
access, based on
multi-LLM token
consumption**



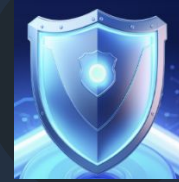
**[Enterprise] Token-Based Rate
Limiting**



**[Enterprise] Azure
Content Safety
Introspection**

Multitude of New Functionality in 3.7

The first of many content safety/moderation services that we will support. Works on all providers and all models, not just Azure



[Enterprise] Azure Content Safety Introspection

AI / LLM Protection Features

How do we help solve the latest IT compliance problem?

01

Regular-expression
guarding against
specific phrases, or
patterns

02

Azure Content Safety,
measure and stops hate,
specific topic blocklists,
and other abuse

03

Token Rate Limiting,
prevents misuse and
overuse of models.
Protects safety and cost.



Powering the API world

Thank you!

[Konghq.com](https://konghq.com)

Kong Inc.

contact@konghq.com